Analyse exploratoire par k-Coclustering avec Khiops CoViz

Bruno Guerraz*, Marc Boullé*, Dominique Gay*,**,
Vincent Lemaire*, Fabrice Clérot*

*Orange Labs
**Laboratoire d'Informatique et de Mathématiques
Université de La Réunion

Résumé. En analyse exploratoire, l'identification et la visualisation des interactions entre variables dans les grandes bases de données est un défi (Dhillon et al., 2003; Kolda et Sun, 2008). Nous présentons Khiops CoViz, un outil qui permet d'explorer par visualisation les relations importantes entre deux (ou plusieurs) variables, qu'elles soient catégorielles et/ou numériques. La visualisation d'un résultat de coclustering de variables prend la forme d'une grille (ou matrice) dont les dimensions sont partitionnées: les variables catégorielles sont partitionnées en clusters et les variables numériques en intervalles. L'outil permet plusieurs variantes de visualisations à différentes échelles de la grille au moyen de plusieurs critères d'intérêt révélant diverses facettes des relations entre les variables.

1 Khiops CoViz : Visualisation des modèles en grille

Khiops CoViz, développée en Flex, est la brique logicielle de visualisation de Khiops Coclustering (KHC) [1]. Étant données, deux (ou plus) variables catégorielles ou numériques, KHC réalise un partitionnement simultané des variables : les valeurs de variables catégorielles sont groupées en clusters et les variables numériques sont partitionnées en intervalles – ce qui revient à un problème de coclustering. Le produit des partitions uni-variées forme une partition multivariée de l'espace de représentation, i.e., une grille ou matrice de cellules et il représente aussi un estimateur de densité jointe des variables. Afin de choisir la "meilleure" grille M^* (connaissant les données) de l'espace de modèles $\mathcal{M}$, nous exploitons une approche Bayésienne dite Maximum A Posteriori (MAP). KHC explore l'espace de modèles en minimisant un critère Bayésien, appelé $cost$, qui réalise un compromis entre la précision et la robustesse du modèle :

$$cost(M) = -\log(\underbrace{p(M \mid D)}_{\text{posterior}}) \propto -\log(\underbrace{p(M)}_{\text{prior}} \times \underbrace{p(D \mid M)}_{\text{vraisemblance}}) \quad (1)$$

KHC construit aussi une hiérarchie des parties de chaque dimension (i.e., clusters ou intervalles adjacents) en utilisant un stratégie agglomérative ascendante, en partant de M^*, la grille optimale résultant de la procédure d'optimisation, jusqu'à $M_\emptyset$, le modèle nul, i.e., la grille (unicellulaire) où aucune dimension n'est partitionnée. Les hiérarchies sont construites en fusionnant

1. `http://www.khiops.com` – Pour plus de détails sur l'implémentation de KHC, voir Boullé (2011)